Deep Learning and Distributed Data Storage System in Identity Recognition and Account Security

Xiangying Wei[1]*, Wei Feng[23]*, Shuyuan Wan[1], Jie Xu[1], Junhao Liu[1], Qujiang Lei[1†], Weijun Wang[1]
[1]Guangzhou Institute of Advanced Technology, Chinese Academy of Sciences, Guangzhou, China
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[3]University of Chinese Academy of Sciences
*Co-First Authors
†Corresponding author
e-mail: qj.lei@aaia-ai.org

***Abstract*—Because of the growth of the Internet, everyone uses a lot of applications and web services. Each user needs an account when using a network service or application. Companies and businesses also link user accounts with user information, resources, and permissions. Therefore, account security is of paramount importance. This article discusses how to protect users' accounts and prevent account security problems caused by the disclosure of some of their personal information. A new method of storing authorized information is proposed to improve the security of user account.**

Keywords-decentralized system; multi-factor; distributed system; asymmetric encryption; deep learning.

I. INTRODUCTION

Mobile computing and mobile application are becoming more and more popular among users. However, the growth of mobile technology also leads to the growth of cyber fraud, such as identity theft or leaking of information. According to *CHINA-DAILY*, users lost Four billion dollars because of identity theft in 2019, and are facing a series of serious follow-up problem for a long time.

Identity theft is costly related to the users' identity authentication. If the system can accurately identify the user's identity when they try to log in to their mobile account, the risk of losing assets will be reduced. This paper discusses the existing authentication methods and puts forward the idea of using distributed structure and deep learning knowledge to better protect user identity.

The existing identity authentication system is centralized structure system and our idea is the decentralized system.

As an old says: "don't put all eggs in one basket." Unlike the traditional centralized system, this paper purposes a different structure of the network communication system, the distributed system. Our idea is using the decentralized structure system to split the risks. The distributed system authentication will avoid the risk of the traditional centralized system.

II. FACTOR AUTHENTICATION AND ITS PROBLEM

Currently, the identification mechanism used in traditional authentication is "username plus the password". Since the username is easy to see, this mechanism is one- factor authentication, which only relies on the password to ensuring its security.

The average user USES two types of passwords: static and dynamic. Static passwords are usually set by users themselves. For security, users are required to set complex passwords with upper- and lower-case letters or even special characters. Static passwords are changed only if the user wants to change them. The dynamic password is a password provided by the server and is a one-time password. It is usually sent to the user's mobile device or the server to give a user a password machine that changes over time (more commonly used in banking).

There are many drawbacks to static passwords. For example, the crime can attack the online server's security hole to steal the password of its users. Besides, the user is easily seen when entering the password. The password can also leak when the network is off line.

In theory, the one-time dynamic password is more secure than the static password, but it still can be stolen or even taken by force. Especially with the appearance of pseudo base station, the security of one-time dynamic password becomes very low. Passwords are a kind of single-factor authentication, once leaked, users' information and property will be put at risk. Traditional single-factor authentication has the lowest level of security, even if the user has a long and complex password.

A. Possible Way Losing Password

- Virus

Viruses are a common means of obtaining information illegally. In particular, many viruses can spread quickly from hidden files, or even from an account after it has been hacked. Its invisibility and variability make it difficult to prevent it from breaking into an account.

- Hidden security holes in the operating system

Every operating system is not perfect, there are some loopholes and defects that were made during the development and designing stage of the operating system.

Especially the open-source operating system, like Android. The system could crash during the file uploading or app installs. The system crashing could lead the leaking of important information of the user [1].

• Credential Stuffing

By collecting the leaked user and password information, the hacker can generate corresponding dictionary tables and obtain a series of user information that can be logged in after trying to log in other websites in batches. Because reuse of passwords is so common, hackers can steal corporate information by using personal accounts to try to log into employees' work accounts.

• Phishing emails

Hackers use disguised emails to trick employees into returning account Numbers, passwords and other information to designated recipients, or guide employees to enter account Numbers, passwords and other information on a special webpage. According to the FBI's 2017 Internet Crime Report, released on May 7, online fraud in the United States caused $1.4 billion in losses in 2017, with phishing emails leading the way with $676 million.

• Take by force

Sometimes the criminals do not have any knowledge of computing. They can just take the user by force. When forced to surrender all the accounts and password, they can only depend on the system to protect their accounts.

• Brute force

When a user uses a static password, it is effective to try all possible brute force methods. Because brute force cracking takes a lot of time, dynamic password is more secure in this respect. However, as a single factor authentication method, it may still be decided by luck.

III. EXISTING SOLUTION

A. Multi-factor Identity Authentication

Identity factors can also be extended to other areas, such as authentication for location and authentication for time [2].

• Time certification.

The authentication is largely used in the bank and only allows the user to use their credit or debit card at a certain time of the day. For example, the user set the time certification for his or her debit card, so that this card can only be used to withdraw money at 11:00 am to 11:05 am. This will effectively enhance security, but at the same time it is a restriction on user usage.

• Location certification.

This method is related to the mobile IP address and it is largely used by mobile application company. A single account cannot be logged in at two different IP addresses. Once that happens, the system will automatically lock the account and send a confirmation to the user's email. On a bank account, it is often restricted to where the card can be used, with the option of a "foreign lock". This method can guarantee the security of the bank account if the user's bank card is lost abroad.

• Multi-device certification.

This credential is widely used for game account security. When an account logs in, it needs to be allowed on another mobile device. This way can prevent account passwords being stolen and causing huge losses.

• Network signal credentials.

Such credentials rely on limiting a fixed type of network signal for security purposes. For example, not allow to login under the condition of WIFI signal. This way can be used as security credentials on the one hand, but also can prevent pseudo base station, fake WIFI hot spot and fake hot spot registration page from stealing user account password information

Authentication for location is based on the mobile IP and authentication for time is based on the timing. Other identity factors can also confirm user identities, for example, the number of transactions and quotas in the e- commerce system can also confirm the user's identity. The more identity factors system has the more information that can be used to confirm the identity of the user. Extending the user's identity authentication and using multiple factors for identity information authentication can enhance the security of the system.

B. Deep Learning in Multi-factor Authentication

• Deep learning and BP neural network

The BP (Back Propagation) neural network is a feedforward type network with the multi-layer structure, which relies on the backpropagation learning [3].

The BP neural network usually consists of multiple layers of neurons. It gives the input to the top layer, and each neuron of this layer will study the input and give output to the next layer. The input data is allocated to the various neuron nodes [4].

•Activation function Sigmoid used by BP network

BP neural network adopts nonlinear transformation function sigmoid function. Its characteristic is that the function itself and its derivative are continuous, so it is very convenient to deal with it. There are two kinds of s-functions: unipolar and bipolar [5].

Unipolar S-type functions (Fig.1) $f: \mathbb{R} \rightarrow (0,1), s.t.$

$$f(x) = \frac{1}{1+e^{-x}}$$

Bipolar S-type functions (Fig.2) $t: \mathbb{R} \rightarrow (-1,1), s.t.$

$$t(x) = \frac{1-e^{-x}}{1+e^{-x}}$$

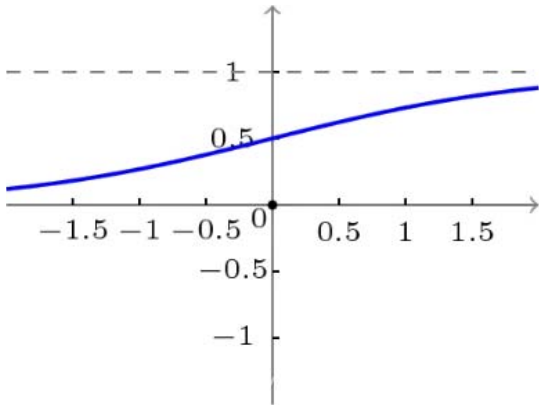

Figure 1. Unipolar S-type functions

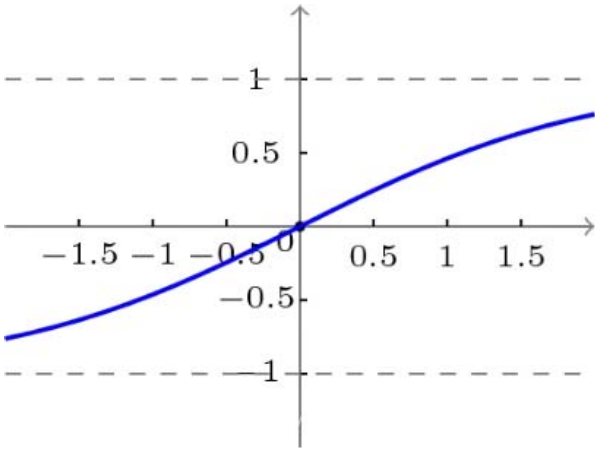

Figure 2. Bipolar S-type functions

•Cost function of BP network

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} log(h_\theta(x^{(i)}))_k + \left(1-y_k^{(i)}\right) log\left(1-((h_\theta(x^{(i)}))_k\right)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(\theta_{ji}^{l})^2$$

K: represents the number of types in multiple classifications.
m: The number of training samples.
$h_\theta(x)$: The y value predicted by the parameter and x;
y: The y value in the original training sample is the standard answer
i: the ith sample
$(h_\theta(x^{(i)})_k$: is the Kth member of the k-dimension vector.
$\frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(\theta_{ji}^{l})^2$: will all Θ accumulation (in addition to bias, the unit above Θ).

•Process of BP network

The learning process of BP neural network consists of the forward propagation of information and the error back propagation [6].

First, input information is input to the input layer, which processes the information and transmits it to the neurons in the middle layer or hidden layer. The middle layer can be designed as single or multiple hidden layers [7].

Then, the hidden layer processes the information and passes it to the output layer. After the final processing of the output layer, the result of a forward propagation is output.

When the actual output does not match the expected output, the error back-propagation begins. The error propagates from the output layer to the input layer to correct the weights of each layer.

Through continuous forward propagation and back propagation, the pre-set target accuracy can be obtained by adjusting the weight of each layer to forward propagation, or stop training after a specified number of times (Fig.3). Theoretically, an effective neural network model can be obtained after a large number of iterations [8].

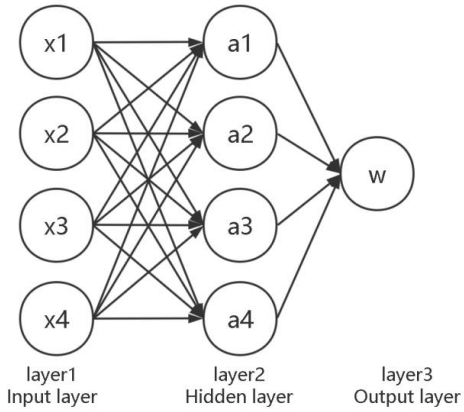

Figure 3. Structure of the BP nature network

• BP neural network in multi-factor identity authentication.

BP neural network can meet the needs of multi-factor authentication.

Each authentication factor can be the expected output of each layer. When users give their login information, the information will be converted to data and distributed in each neuron in the top layer. During the learning stage, the neurons apply the match rules and find the best match to the input from the match-base. If the input of the user match with the expected result, the input will be taken to the next layer for another round of learning. The button layer will give the studied result. If this result match with the expects result, the system will premise the user to log in.

During the process of deep learning, new judgment rules can be added to the matching base. The machine can adapt

the latest rule and algorithm such as voice, signing, gesture or even a smile. The more and more complex the factors are, the more accurate the system will be.

C. Structure of the Multi-factor Identity Authentication

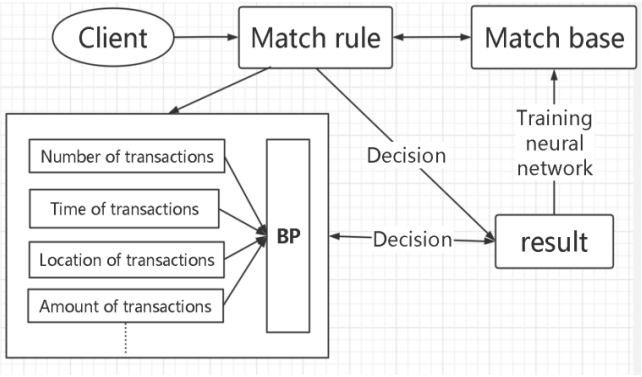

Figure 4. Structure of the multi-factor identity authentication.

• Client

The Client side is the UI for users to input their authorization information. (Fig.4)

• Match rule

The Match rule contain the authorization information database that find the match with the input information from user. For instance, in the system above the match rule above contains four rules for user to match: number of transitions, time if the transition, location and amount. Only if the user's input information match with all correct rules, the system will let the user log in to the system. (Fig.4)

• Match base

Match base perform the authorization step, where the user's input information will compare with the match rule. This is where the deep learning going on. For instance, in the BP neural network system, the classifier of CNN and RNN will recognized the user input information and give the result. If the result match with the match rule, the system will give the full access to the user. (Fig.4)

D. Deep Learning in Identity Recognition

• Biometrics technology

Biometric technology is using physical characteristics and behavioral characteristics of the human body to confirm the user's identity. Since each person's biometrics are different, using this method to verify the user can ensure accurate user ID [1].

The biometric identity authentication method is relatively reliable, and it is the most advanced authentication method.

The biometrics feature we use in our system includes facial, and fingerprint.

• Facial

The facial recognition involves with the CNN. first the user input their front face in the match rule. The system will scan the image and give a matrix of 128*128*3. When the user wants to log into the system. He or she will scan their face in the client side, and system will give a 128*128*3 matrix for the match base to match. The test loss is 0.080000943850. The test accuracy can reach to 0.9766 (Fig.5).

• Fingerprint

The fingerprint recognition involves with the CNN. first the user input their fingerprint in the match rule. The system will scan the image and give a matrix of 128*128*3.

When the user wants to log into the system. He or she will scan their fingerprint in the client side, and system will give a 128*128*3 matrix for the match base to match. The test loss is 0.05652410360. The test accuracy can reach 0.9824 (Fig.6).

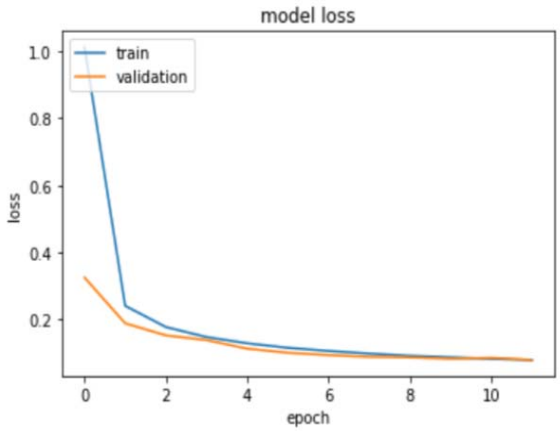

Figure 5. The test result of facial recognition

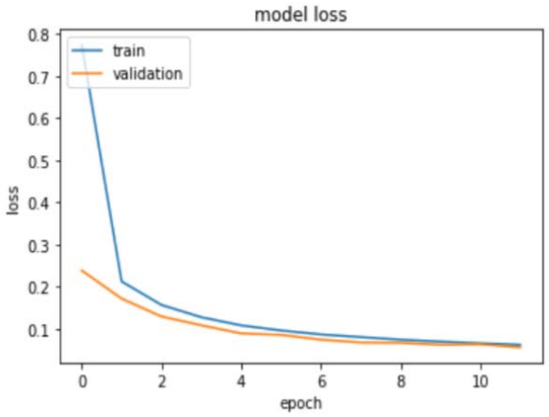

Figure 6. Test result of fingerprint recognition.

•CNN

As one of the representative algorithms of deep learning, Convolutional Neural Networks (CNN) is a kind of Feedforward Neural Networks that contains Convolutional computing and has a deep structure. CNN is generally composed of three layers [9]:

The convolutional layer:

The convolution theorem says that the Fourier transform of the convolution is the product of the Fourier transforms of the functions. That is, the convolution of one domain is the product of another domain [10].

Generally, there are many different convolution layers in a CNN, and each convolution layer has its corresponding kernel (Fig.7). Convolution kernel K is a square with a fixed size of f. the input data is scanned by setting the strip s for setting the scanning step size and padding p for filling the input image. Each scan area of input data is convoluted with the kernel [11].

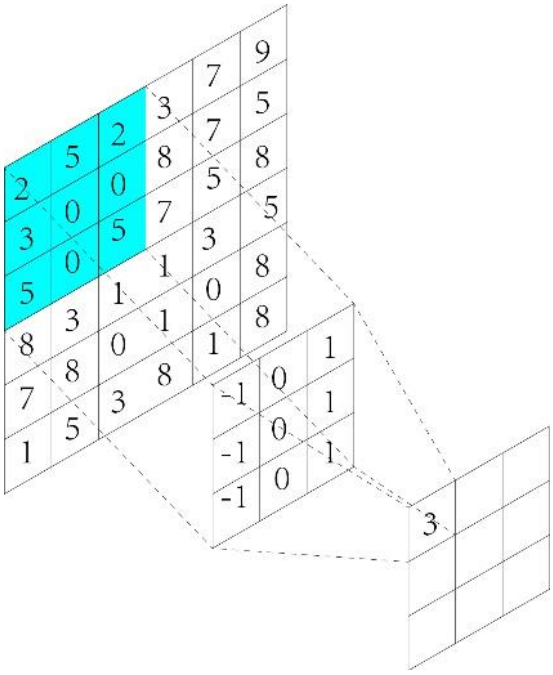

Figure 7. The process of Convolution

The convolutional product C is a mapping of : $M \rightarrow R$

$$C = \sum_{i=1}^{f-1} \sum_{j=1}^{f-1} M_{ij} \times K_{ij}$$

K is the kernel.

f is the size of the kernel.

M is the scanned area matrix of the input data.

The Pooling layer (POOL):

The pooling layer also has a number of kernels for scanning input data. At the same time, the eigenvalues of the set region are extracted to remove invalid or relatively irrelevant parameters, which will prevent overfitting. Generally, there are two pooling methods. One is Max pooling, which extracts the maximum value of the scanned area as the output value. The other is the average pool, which calculates the average of all values in the pool area as the output [12]. Figure Fig.8 is the result of 2x2 pooling of a 4X4 matrix.

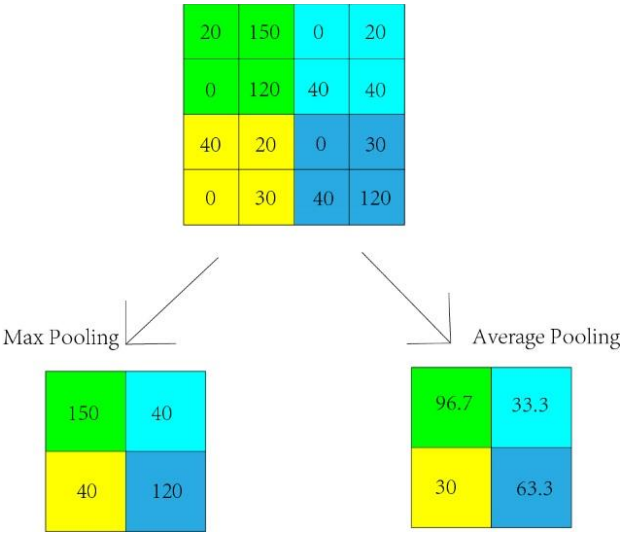

Figure 8. The two types of Pooling

The Fully Connected layer (FC):

The most common network layer is generally used as the last few steps of CNN network. Because each input of each layer links every output of the previous layer, it is called full connection layer [13].

The CNN structure we use here is a VGG 11 structure. There are eight convolutional layer and three full connected layers, each conversational layer contains kernel size of three, pooling size of two and one stride (Fig.9).

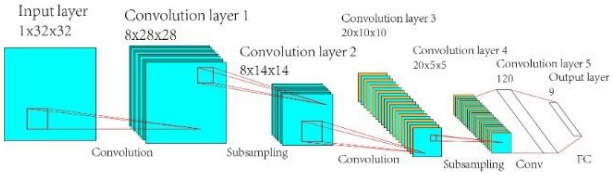

Figure 9. The structure of CNN

•Drawback

Both Multi-factor and biometrics technology are belonging to the centralized system. Therefore, they are born with drawbacks of a centralized system. No matter how complex authentication algorithm used, there always a centralized server to decide whether to permit the user to log in. *However*, there is always a possibility that this server can to be cracked. In the centralized system, the center is holding all the data and authorization for the users. The crashing of

the center is going to put all system in danger. In this case, it refers to talking about the leaking of millions of users' information and identities. Because of that those exiting mothed can not 100% protect users' identification. For example, according to CNN (Cable News Network) report, the Equifax's database was cracked at July 29th 2017, 143 million's people are affected.

IV. SOLUTION

Our idea and solution

Instead of letting a centralized system to identify users' identity. Our idea is building a distributed system to let many participate blocks in the system to do that job. This idea builds multi-factor authentication with distributed system, which avoids all the risk that the centralized system has, and the and make the data invincible [14] [15].

•Blockchain

Blockchain system is a decentralized system involved with distributed data storage, point-to-point transmission, consensus mechanism, and encryption algorithm. The most popular thing that relate to the Blockchain is Bitcoin. The advantage of blockchain are [16] [17] [18] [19]:

1) Decentralized. Blockchain system does not rely on additional third-party centralized server or hardware facilities to manage the system. Without central control, each block self-verification, serf-delivery, and self-management.

2) Independence. Based on a consistent protocol, the entire blockchain system does not rely on any management, all blocks can freely verify and exchange data within the system. Every block is independent. Breaking one node does not cause the whole system to crash.

3) Safety. As long as hackers do not control 51% of blocks. They can't decrypt any data, because, in the blockchain system, every information is encrypted and then distributed to every block in the system.

4) Public. The blockchain system is open to the public. Every person can join a blockchain.

•How blockchain authentication works

First, the user creates an account, which means the user creates a new block in the blockchain. User can choose his or her multi-factors log in information and tell the system how to identify his or her identity for the next login. The system will first encrypt that information and then broadcast that to other participate blocks in the blockchain. The other blocks will record part of this encrypted information. When this user wants to log in to the account again, this user's block will ask authorization from all other blocks in this system [18].

For example, there are six blocks in a blockchain system. If the user chose to use password, time and location certifications, every other block will record one certification. The information the other block record is encrypted. If user's block is bock one, block two will record the password; block three will record time; block four will record location; block five will record password again, block six will record location again (Fig.10).

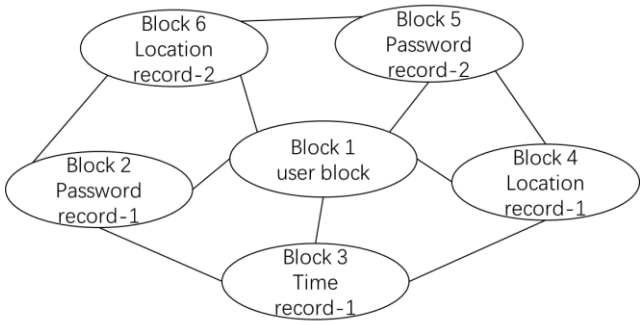

Figure 10. A simple Block chain login system.

When a user attempts to log in to a block again, the first block requests authorization to log in from other blocks in the system (Fig.11).

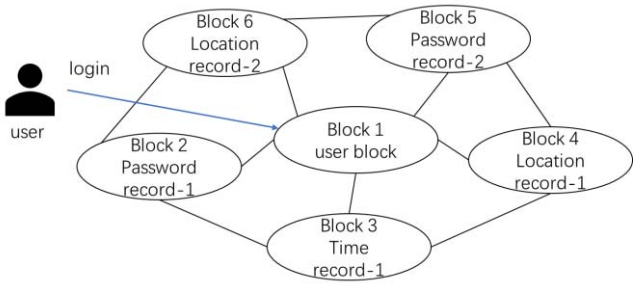

Figure 11. User try to login.

Each block will receive a request from the block 1 (Fig.12).

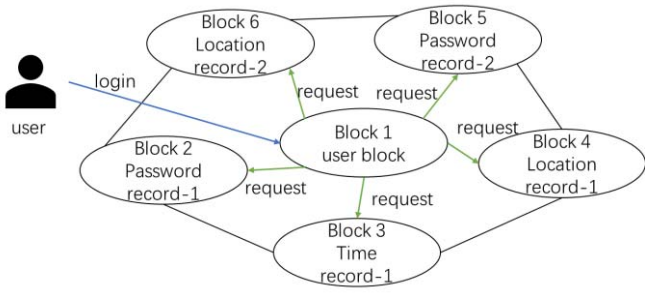

Figure 12. Block 1 send request.

If all blocks are authorized successfully, the login will succeed (Fig.13).

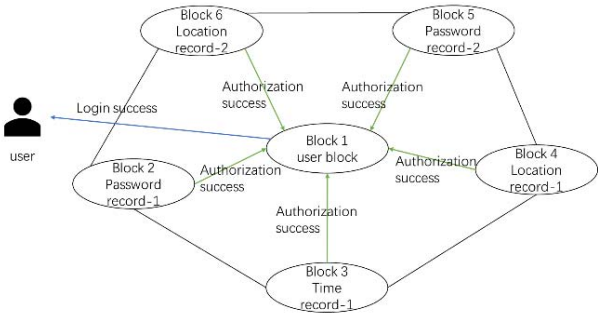

Figure 13. Blocks authorization success.

If one or more blocks deny permission, the login fail (Fig.14).

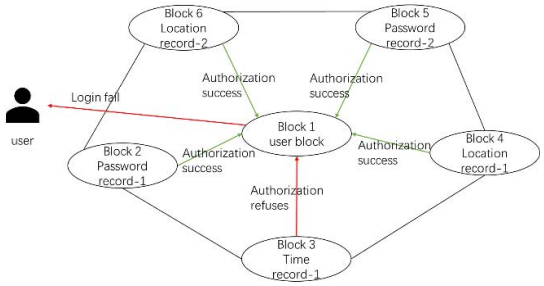

Figure 14. Login fail.

If one of the blocks is hacked. The intruder cannot get all the authentication information, so the hacker cannot log in to the user to get the user information (Fig.15). At the same time, even if a hacker destroys the contents of a single block, there is no data loss because the same information exists in other blocks. To get user information, hackers crack at least 51% of the blocks, which can be a lot of work when the number of blocks associated is large.

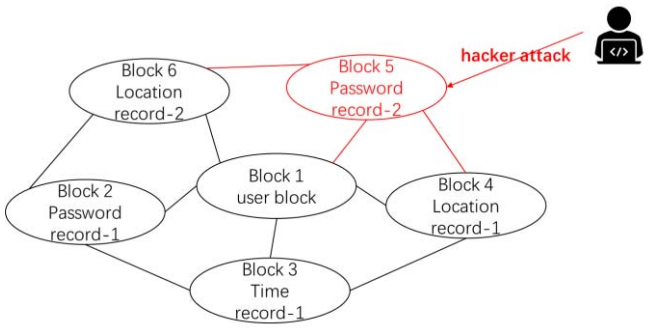

Figure 15. Hacked.

•Multiple supervision system

In this multiple supervision system, Distributed system let other users to inspect one user's identity authentication, but a user's log in information's security is not heavily relied on one user.

•Advantage of using decentralized distributed authentication

Because there is no centralized decision-making system, the decentralized distributed avoids the risk of a centralized system. If a hacker cracks one block, they are only breaking one part of the encrypted information, *which is impossible to be decrypted*. Even if the hacker successfully decrypts the information, the hacker still cannot log in to the user's block because the hacker does not have all the login information and will get rejected by other blocks.

In order to get all login information, the hacker will have to crack all the blocks and decrypt their information in the distributed. Since none of the blocks is core, there are even a large number of blocks that have duplicate information with other blocks. A broken block does not cause data loss or loss of credentials.

•Asymmetric encryption

Why the Distributed is so confident the hacker will not decrypt the information of one block, because of the Asymmetric encryption. Unlike the Symmetric encryption only uses one key. Asymmetric encryption uses a public key to encrypt data and private key to decrypt data. The public key is in the system, while the private key is holding by users. If the hackers get the public key, they are not able to decrypt the data without the private key.

V. CONCLUSION

This paper proposes a discrete distributed system, which establishes multi-factor authentication to protect the user's identity information and account information. Compared with traditional authentication, multi-factor authentication has the advantages of security, independence, encryption, distribution, cooperation and stability in distributed system. It can effectively prevent the single system of centralized system from being broken and causing great harm. Distributed systems allow each user to store and protect data. The collapse of a single block does not result in the collapse of the entire system. The identity of the user and important information in the system will be secure.

• Challenges

Distributed systems require ultimate network environment.

Theoretically, the more blocks in the system, the safer the system is. However, this requires that the blocks are all stable in the system. When a block is making the decision, any disconnection will cause misunderstanding and possibly lead to making the wrong decision in the whole picture.

Mobile computing Distributed system also needs huge bandwidth, because when a user is logging in, a single block will need to send and receive information from every block in the system.

Every block in the distributed needs a large mobile database. Not only does the database need to record login information from all other blocks, but also the database needs to store the decrypt method to decrypt information from other blocks.

Every block requires a high-performance central processing unit in order the decrypt the information. Because even if the block already stores the method to decrypt the

information, some of the decrypt methods require a large amount of computing.

For example, there is a Random hash function. Any length of data can be converted into a 256-bit code of 0 and 1. Any small change in the source data will cause a huge change in the encrypted data. If one wants to get a meaningful code out if it, there are no other ways but “Brute-force attack”, which cannot be done by the current central processing units in the mobile device.

• Possible way to solve challenges

The possible solution is 5G. From 1G (0.9GHz) to 4G (above 1.8GHz), the frequency of wireless electromagnetic wave we use is increasing. The main advantage of 5G networks is that the data transmission rate of 5G is much higher than the previous cellular network.

The transmission rate is up to 10 Gbit/s, which is faster than the current home wired Internet and 100 times faster than the previous 4G LTE cellular network. Another advantage is lower network latency and faster response time; 5G has network latency of less than 1 millisecond, and 4G is 30-70 milliseconds. 5G has a large network capacity, its network capacity is 1000 times that of 4G, and its peak rate is 10gbps-20gbps, which means that it adopts a higher frequency band, which can provide a connection capacity of 100 billion devices' connection. At the same time, it can maintain low power consumption of battery life.

5G can achieve real-time communication, high-speed Internet access, do not worry about losing contact with the land, the network is like air, no blind area, no dead corner, full coverage. Stable, high spend, high performance, faster response time, easy to carry with you, cloud database, those are the main factor for Distributed authentication.

ACKNOWLEDGEMENT

This research is supported by the Guangdong Province Key Field R&D Program Project (Grant No. 2020B090925002), the Major Projects of Guangzhou City of China (Grant No. 201907010012).

REFERENCES

[1] L. G. R.P. Matei, "Design methods for CNN spatial filters with circular symmetry," Neural Network Applications in Electrical Engineering 2004 7th, pp. 103-108, 2004.

[2] G. E. Hagopian, "Analog cellular image sensor processor (CISP)," Circuits and Systems 2000. Proceedings of the 43rd IEEE Midwest Symposium on, vol. 2, pp. 620-623, 2000.

[3] D. E. G. E. H. a. R. J. W. Rumelhart, "Learning representations by back-propagating errors," Nature 323.6088, pp. 696-699, 1988.

[4] Y. Lecun, "Handwritten Digit Recognition with a Back-Propagation Network.," neural information processing systems, pp. 396-404, 1989.

[5] J. Mount., "The equivalence of logistic regression and maximum entropymodels[J]," 2011.

[6] P. S. T. Szabo, "Morphological and wave segmentation by cellular neural networks," Neural Network Applications in Electrical Engineering 2004 7th Seminar on, pp. 109-112.

[7] Y. a. L. B. Lecun, "Gradient-based learning applied to document recognition," Proceedings of the IEEE 86.11, pp. 2278-2324, 1998.

[8] A. a. B. N. Van Ooyen, "Improving the convergence of the back-propagation algorithm.," Neural Networks 5.3, pp. 465-471, 1992.

[9] L. Yann, "Backpropagation applied to handwritten zip code recognition.," Neural Computation 1.4, pp. 541-551, 1989.

[10] L. G. C. M. S. C. T. A. B. A. Lawrence S, "Face Recognition: A Convolutional Neural-Network Approach," IEEE Transactions on Neural Network, vol. 8, no. 1, pp. 98-113, 1997.

[11] J. Bouvrie, "Notes on Convolutional Neural Networks," 2006.

[12] a. B. W. R. Francis E. Raymond, "Fractional Max-Pooling," Theoretical Economics Letters, pp. 225-237, 2 5 2015.

[13] P. G. K. I. Liangchieh C, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J].," Computer ence, pp. 357-361, 2014.

[14] D. K. S. L. Y.M. Qin, "China cold-chain logistics development report (2014)," in, Beijing: China Fortune Press, pp. 116-117, 2014.

[15] A. T. G. Foroglou, "Further applications of the blockchain," Columbia University PhD in Sustainable Development 10 Year Anniversary Conference, p. 2014.

[16] V. Gatteschi, F. Lamberti, C. Demartini, C. Pranteda and V. (. F. 2. Santamaría, "Blockchain and Smart Contracts for Insurance: Is the Technology Mature Enough?". Future Internet. 10 (2): 20. doi:10.3390/fi10020020.

[17] a. S. G. B. K. Kotobi, "Secure Blockchains for Dynamic Spectrum Access: A Decentralized Database in Moving Cognitive Radio Networks Enhances Security and User Access," IEEE Vehicular Technology Magazine, 2018. DOI: 10.1109/MVT.2017. 2740458, 2018.

[18] M. Toorani, "A. Beheshti. 2008. SSMS - A secure SMS messaging protocol for the m-payment systems.," IEEE Vehicular Technology Magazine 2008. DOI: 10.1109/ISCC.2008.4625610, 2008.

[19] K. Saito and H. Yamada, "What's So Different about Blockchain? Blockchain is a Probabilistic State Machine.," IEEE 36th International Conference on Distributed Computing Systems Workshops. Nara, Nara, Japan: IEEE. doi:10.1109/ICDCSW.2016.28., 6 2016.